Philosophical Problems Paper III
By Sophie Niegsch and Ludovica Schaerf

The Issue of Equivalence between Black Box and White Box Algorithms

An Analysis of the Ontological Value of Interpretability and Explainability

Introduction

Nowadays, Artificial Intelligence' (AI) procedures that classify as 'black boxes' are at the basis of most of the algorithms used by our devices. Such procedures, albeit classifying the data with high precision and accuracy, consist of a decision-making process that is - almost always - uninterpretable and unexplainable. This issue has opened the debate both in industries and elsewhere of whether it is safe, or at least ethical, to rely so extensively on algorithms that lack in accountability (Molnar, Christoph., 2019). This debate is particularly central when concerning high-risk situations that have not been fully understood by previous researches (Doshi-Velez and Kim 2017). A controversial instance of this are AI self driving cars: if we consider a scenario in which the car is faced with the decision of either killing the people crossing the road or derailing and killing the passengers of the car, it is retrospectively essential to be able to account for the decision-making process of the AI.

In contribution to this debate, this essay explores the notion of equivalence of algorithms, focusing on whether their interpretability and explainability are aspects that must be taken into account when making judgements on equivalence.
In the analysis, the concept of equivalence is explained both philosophically and in computer science terms, and translated into the problem of reducibility of an algorithm into the other and into the Halting problem. Finally, the paper argues in favour of a conception of an output as an entity that expresses the concept of the algorithm it has been produced by, entailing that the equivalence in outputs of two algorithms fails when the concept of one cannot be recreated.

Assumptions and Limitations

This paper exploits the concept of equivalence both in computer science and philosophy, assuming that the entities, as they are defined in the two fields, are reducible to one another. This broader assumption entails that the proof for undecidability in computer science applies to philosophy, which implies that equivalence as a philosophical entity is also undecidable.

To support this last point, the essay accepts the view of token physicalism and the idea that the bridge law between equivalence in philosophy and in computer science pertains to the events, but there are some properties which do not reduce to physical ones. For instance, equivalence in computer science is a Turing Complete problem that can be reduced to all other problems in the set; in philosophy, however, such a reduction is impossible. Following

again from token physicalism, the essay assumes the possibility for an output to contain non-physical properties.

The discussion on causation and correlation accepts Kant's view of causation as a mental structure in the human mind - and supports the view that machines alone are unable to construct causal links and only extrapolate correlations.
Lastly, it is important to keep in mind that equivalence is bound by various contextual factors, such as linguistic constructs in which, under certain circumstances, logical equivalences are epistemological constructions.

Definitions

The section that follows builds the argumentation by means of a serie of definitions in the two disciplines that are brought together at the end of each subsection.

Equivalence

Philosophy

Equivalence has been defined in the field of logic as follows:

> 'Equivalence refers to propositions or formulas that share the same logical meaning. Equivalent propositions or formulas have the same truth value regardless of the valuation of their terms.' (Philosophy Index)

Which means that two propositions are equivalent if they are evaluated either both as true or both as false. However, real life properties and entities cannot be fully expressed by the dual state of logic, as for example in the case of two trees, the equivalence between the two is impossible to express in a finite number of propositions.

A broader definition is therefore required: in general philosophical terms, equivalence is concerned with the question of whether the structuring of the necessary relations of similarity and difference that construct the identity of the entities translate between epistemological constructions through functions of reflexivity, symmetry and transitivity. The identity of an entity or object concerns the properties by which one thing is that one thing but no other thing. In other words, two entities are equivalent if one can be reduced through reflexivity, symmetry and transitivity into the other completely.

This begs the question that will be addressed later: can we regard two things the same if we don't know what is going on in at least one of the two? Or to rephrase: can we translate an

identity into a different world, a different model, if we don't understand the epistemological construction that guides it?

Computer Science

The concept of equivalence between two algorithms[1] has been defined in computer science on many different levels. In terms of behavioural equivalence - the one this essay focuses on - it is often phrased in the following two ways:

> 'Loosely speaking, two expressions M and M' of a programming language are contextually equivalent if any occurrences of M and M' in complete programs can be interchanged without affecting the results of executing the programs' (Pitts, Andrew M., 1997)
>
> 'A binary relation connecting algorithms of a fixed type and expressing the fact that any two algorithms thus connected yield the same results if fed with the same given type of inputs (and may also yield some additional information as regards the computations thus performed — the so-called history of the computation).' (Encyclopedia of Mathematics)

From these definitions, it becomes clear how computer science frames equality in terms of equality of outputs for the full range of possible inputs, but the second definition also demonstrates how the 'history of the computation' - almost as a side effect of the procedure - may be influential in the equality.

In addition, the problem of equivalence has been precisely defined in terms of equivalence of Turing Machines (TM). Once two algorithms have been rewritten into TMs, if another TM (denoted in the sequel as EQ_TM), that determines whether two TMs are the same, accepts, then the two algorithms are the same. However, with the exception of extremely simple cases, it is easy to prove that the EQ_TM is not guaranteed to halt in a finite amount of time. This phenomenon proves that equivalence of two machines is undecidable: it is a problem that is Turing Complete[2], reducible to the Halting problem.

Therefore, the problem phrased in computer science terms becomes: does EQ_TM, when it receives as input an interpretable program and a non-interpretable one, always refute? Or is the problem still undecidable even in the case in which one is a black box and one is not? In

[1] By algorithm we mean very broadly any series of steps that are unambiguous and computable
[2] The set of problems that are undecidable and reducible the one into the other

other words, is the output of a black box ontologically different from the one of a white box, case in which the TM would halt and refute?

Synthesis

As in philosophy, in order to be able to translate between one epistemological construct to another, the concept of the algorithm has to be understood for it to be translated, because otherwise there would be no guarantee that the algorithm is fully translated. Similarly, loosely mathematically speaking, the proof (here adopted in a broad sense, analogous to the concept of algorithm) must be known in the output to provide a guarantee of the validity of it. In fact, given an algorithm that maps an input *i* to an output *o*, to prove that *o* is a valid deduction, the information given, implicitly, by *o* must contain the reason (the proof) why the deduction from *i* got to *o*.

If we accept that the output incorporates the concept of the algorithm, the EQ_TM would always reject (meaning that the two algorithms are not equivalent) when one of the outputs contains such a concept and in the other it is unknown. Generally, even accepting this concept of output, the problem of equivalence remains undecidable when both concepts are given and a TM that judges on equality needs to be applied to them as well.

Black Box and White Box

Philosophy

At this point it is important to explain what exactly does the essay refer to with the term ‘black box’.

Metaphorically introduced in the conception of the human mind and behaviours, the notion of the black-box referred to the hidden capacity of the mechanisms; in which input garned output, but the how, whats and whys in between were matters of mere speculation.
White box mechanisms on the other hand allow for an accessible and clear insight into the thinking and steps followed to reach the output, putting black box mechanisms in a difficult position regarding the investigation of causal or mere correlated functions.
A deduction made by a black-box mechanism therefore lacks of:

- explainability: the capacity to understand the reasoning behind the deduction (Rao, Anand., 2018)

- interpretability: the ability to comprehend what the mechanism did (Gilpin, Leilani H., et al., 2018)

Computer Science

Within the field of Artificial Intelligence, the focus has generally shifted to ‘black box’ procedures. There no longer is a precompiled interpretation of the parameters and data that are fed into the program, that learns under supervision of an external agent. The data and parameters are now fed to a machine that, through unsupervised deep-learning, programs itself by positive and negative reinforcement feedback or backpropagation; re-weighting the end-values to reconfigure its algorithm - and thereby ontological output possibilities - to adapt to solution processes.

Programming entailed an epistemology that was demarcated and pre-figured in thinking; narrowly outlining the possibilities of ontological action in the space. Deep-learning mechanisms evade the necessity of a supervised learning direction that informs their own decisions.

Example: Neural Nets

The most important examples of such procedures are neural networks (deep learning). These algorithms are currently giving promising results in the accuracy and precision of their predictions. However, the multi-layered nature of the process together with the convergence of a level with many nodes to one with fewer, causes the model to be un-invertible, and therefore not interpretable nor explainable.

A more precise explanation of a system of the type of neural nets is contained in Appendix 1.

Reduction

The problem that these definitions originate is again the following: How does the output of a black box reduce the one of a white box mechanism? Can equivalences be drawn through translatory nature?

As seen in the definition of equivalence, the two outputs ought to be considered different because they contain (or not) the concept of the algorithm. To support this point, this paragraph shows how the output of a white box is not reducible to the one of a black box nor viceversa.

Fodor disputes the topic of the reduction of psychological events or situations into neurobiological phenomena, emphasizing that even if there were token equivalence between the two it would “not follow that the natural kind predicates of psychology are coextensive

with the natural kind predicates of any other discipline" (such as neurobiology or neurophysiology) (Fodor, 1974, 105). The relationship between the mind and computer science, or programming in particular, has to be taken into account when dealing with neural networks, as they mimic the properties of biological neuronal networks in the brain, and, likewise, they should reproduce the hidden causation that us humans construct between events. However, unless explicitly programmed like it happens in white boxes, a black box algorithm is unable to understand such relations and distinguish them from correlations.

Considering this aspect, a prediction that is based on the awareness of causal links is ontologically non-reducible into one that is not, as the former contains non-physical properties that the latter doesn't.

For other irreducible aspects of a human-programmed algorithm into a black box one, more information is contained in Appendix 2.

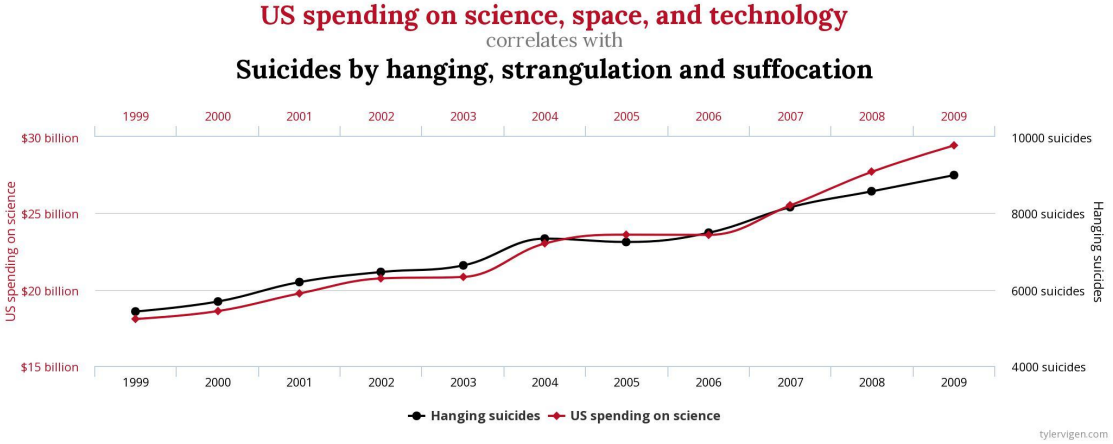
Causation

Philosophy

As discussed in the Assumptions paragraph, this paper adopts the Kantian view of causality. Immanuel Kant refutes Hume's view on causation - which had deprived causation of any ontological existence, lowering it to mere habit - , by placing the causal link among the structural categories of the mind. In this way, Kant gave an account of causality that merges Hume's consideration of causation as not a phenomenon, with the urge to give it a scientific validity (Kant, Immanuel, 1999).

Computer Science

In machines this ability is not similarly innate and it has to be explicitly programmed by the coder. In black-box algorithms, in which it is not provided by the human, there is only an extrapolation of correlations from the data. This introduces a high risk that the biases within



the data set appear in the output. For instance, a neural network that has been given the data on US spending on science, space and technology and the data on suicides by hanging, strangulation and suffocation would extrapolate a correlation of 99.79% between the two data and deduce causality (Spurious Correlations).
The same mistake would not have been made by a white box and it is plausible to assume that many other such spurious correlations are being created by black boxes when classifying, making the output ontologically different from that of a white box.

Conclusion

In conclusion, this paper has analysed the question of whether a ‘black box’ algorithm can even be equivalent to a ‘white box’ one despite being not interpretable and not explainable. To do so, the concept of equivalence has been defined as the equality of the set of outputs for the whole set of inputs. Based on this definition the question has been reshaped into whether the output of a black box is ontologically different from the one of a white box.

The question has been answered by expanding the output to include the non-physical concept of the algorithm and proving that the EQ_TM would always consider such algorithms different.
Finally, by introducing the concept of reduction, the paper has shown an example of non-physical concept of the white box algorithm that is not implementable by a black box: causation. In fact, a machine that has not been given an interpretation by the human cannot reconstruct causal links, therefore making the qualities of interpretability and explainability ontologically valuable in equivalence.

References

“Algorithms, Equivalence Of.” Algorithms, Equivalence of - Encyclopedia of Mathematics, www.encyclopediaofmath.org/index.php/Algorithms,_equivalence_of.

Doshi-Velez, Finale, and Been Kim. “Towards a rigorous science of interpretable machine learning,” no. Ml: 1–13. http://arxiv.org/abs/1702.08608 (2017).

“Equivalence.” Philosophy Index, www.philosophy-index.com/logic/terms/equivalence.php.

J.A. Fodor.‘Special Sciences (or: the Disunity of Science as a Working Hypothesis)’ Synthese 28 (1974), pp. 97-115.

Gilpin, Leilani H., et al. "Explaining Explanations: An Overview of Interpretability of Machine Learning." 2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA). IEEE, 2018.

Kant, Immanuel. Critique of pure reason. Cambridge university press, 1999.

Molnar, Christoph. “Interpretable Machine Learning.” 2.1 Importance of Interpretability, 12 Apr. 2019, christophm.github.io/interpretable-ml-book/interpretability-importance.html.

Narayanan, Ajit. "Fodor and Pylyshyn on connectionism: an extended review and brief critique." Artificial intelligence review 2.3 (1988): 195-213.

Pitts, Andrew M. "Operationally-based theories of program equivalence." Semantics and Logics of Computation 14 (1997): 241.

Rao, Anand. “What It Means to Open AI's Black Box.” Next In Tech, 2 Aug. 2018, usblogs.pwc.com/emerging-technology/to-open-ai-black-box/.

“15 Insane Things That Correlate With Each Other.” Spurious Correlations, www.tylervigen.com/spurious-correlations.

Lecture Slides on Fodor

Appendix 1

‘Connectionist systems are networks consisting of very large numbers of simple but highly interconnected 'units'. Certain assumptions are generally made about the units and their connections. Typically the units do little more than sum this activity and change their state as a function (usually a threshold function) of this sum. Each connection is allowed to modulate the activity it transmits as a function of an intrinsic (but modifiable) property called its 'weight'. Hence, the activity of an input line is typically some non-linear function of the state of activity of its sources. The behaviour of the network as a whole is a function of the initial state of activation of the units and of the weights of its connections, which serve as its only form of memory’ (Fodor & Pylyshyn, 1988, pp. 196-197).

Appendix 2

There are of course still structural and manifest differences between the mind and the computer, amongst which for instance the compositional difference in multiplicity and quantity of neuronal structures and connections in a composite whole account for possibilities exceeding the direct connectional functions of neuronal networks. One compositional attribute, for example, can be fundamental drives such as nourishment and coitus that are not (necessarily) impulsive in their effect but drive and direct behaviour despite lacking feedback for longer periods in learning progression. Components of communicability arise alongside the question of the ethical concern of accountability as the black box of a human still stands accountable for their output actions (to a certain degree and under the exception of certain circumstances), and at least are (usually) capable of explaining their actions or motivations even if not necessarily wholly or necessarily true. With an inaccessible regression chain exemplifying or elucidating the decision and thinking process of the program, there can be no accountability held; not only towards an object (as a fundamental difference) but also a correlation or thought-association that might have directed towards a negative action or output, ideally capable to be modified.